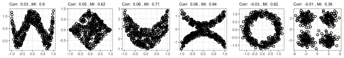
Introduction to Machine Learning

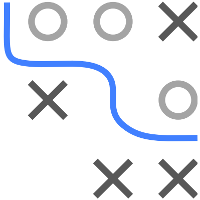
Joint Entropy and Mutual Information II



Learning goals

- Know mutual information as the amount of information of an RV obtained by another
- Know properties of MI

MUTUAL INFORMATION - COROLLARIES



Non-negativity of mutual information: For any two random variables, X, Y, $I(X; Y) \geq 0$, with equality if and only if X and Y are independent.

Proof: $I(X; Y) = D_{KL}(p(x, y) \| p(x)p(y)) \geq 0$, with equality if and only if $p(x, y) = p(x)p(y)$ (i.e., X and Y are independent).

Conditioning reduces entropy (information can't hurt):

$$H(X|Y) \leq H(X),$$

with equality if and only if X and Y are independent.

Proof: $0 \leq I(X; Y) = H(X) - H(X|Y)$
Intuitively, the theorem says that knowing another random variable Y can only reduce the uncertainty in X. Note that this is true only on the average.

Remark: Because $H(X) \geq H(X|Y)$ and $H(X)$ is only bounded from below, $I(X; Y)$ is unbounded from above (lives in all of $\mathbb{R}_0^+$)

MUTUAL INFORMATION - COROLLARIES / 2

Independence bound on entropy: Let $X_1, X_2, \ldots, X_n$ be drawn according to $p\left(x_1, x_2, \ldots, x_n\right)$. Then

$$H\left(X_1, X_2, \ldots, X_n\right) \leq \sum_{i=1}^{n} H\left(X_i\right),$$

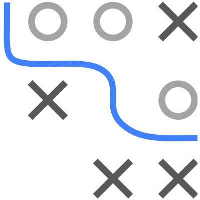
with equality if and only if the X_i are independent.

Proof: With the chain rule for entropies,

$$H\left(X_1, X_2, \ldots, X_n\right) = \sum_{i=1}^{n} H\left(X_i | X_{i-1}, \ldots, X_1\right) \quad \leq \sum_{i=1}^{n} H\left(X_i\right),$$

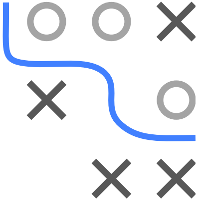
where the inequality follows directly from above. We have equality if and only if X_i is independent of $X_{i-1}, \ldots, X_1$ for all i (i.e., if and only if the X_i 's are independent).

MUTUAL INFORMATION PROPERTIES

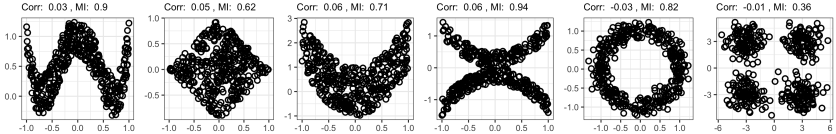


- MI is a measure of the amount of "dependence" between variables. It is zero if and only if the variables are independent.
- On the other hand, if one of the variables is a deterministic function of the other, the mutual information is maximal, i.e. entropy of the first.
- Unlike (Pearson) correlation, mutual information is not limited to real-valued random variables.
- Mutual information can be used to perform **feature selection**. Quite simply, each variable X_i is rated according to $I(X_i; Y)$, this is sometimes called information gain.
- The same principle can also be used in decision trees to select a feature to split on. Splitting on MI/IG is then equivalent to risk reduction with log-loss.
- MI is invariant w.r.t. injective and continuously differentiable reparametrizations

MUTUAL INFORMATION VS. CORRELATION



- If two variables are independent, their correlation is 0.
- However, the reverse is not necessarily true. It is possible for two dependent variables to have 0 correlation because correlation only measures linear dependence.



- The figure above shows various scatterplots where, in each case, the correlation is 0 even though the two variables are strongly dependent, and MI is large.
- Mutual information can therefore be seen as a more general measure of dependence between variables than correlation.

MUTUAL INFORMATION - EXAMPLE

Let X, Y be two correlated Gaussian random variables.
$(X, Y) \sim \mathcal{N}(0, K)$ with correlation ρ and covariance matrix K:

$$K = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

Then $h(X) = h(Y) = \frac{1}{2} \log \left((2\pi e)\sigma^2 \right)$, and
$h(X, Y) = \frac{1}{2} \log \left((2\pi e)^2 |K| \right) = \frac{1}{2} \log \left((2\pi e)^2 \sigma^4 (1 - \rho^2) \right)$, and thus

$$I(X; Y) = h(X) + h(Y) - h(X, Y) = -\frac{1}{2} \log(1 - \rho^2).$$

For $\rho = 0$, X and Y are independent and $I(X; Y) = 0$.
For $\rho = \pm 1$, X and Y are perfectly correlated and $I(X; Y) \to \infty$.

ESTIMATION OF MI

- In practice, estimation of the mutual information

  $$I(X; Y) = H(X) + H(Y) - H(X, Y)$$

  is usually based on the *empirical information*, i.e.,

  $$\hat{I}(X; Y) = \hat{H}(X) + \hat{H}(Y) - \hat{H}(X, Y)$$

Here, we simply plug in the estimates of the empirical distribution $\hat{p}(x), \hat{p}(y), \hat{p}(x, y)$:

$$\hat{H}(X) = -\sum_{x \in \mathcal{X}} \hat{p}(x) \log_2 \hat{p}(x)$$

$$\hat{H}(Y) = -\sum_{y \in \mathcal{Y}} \hat{p}(y) \log_2 \hat{p}(y)$$

$$\hat{H}(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \hat{p}(x, y) \log_2(\hat{p}(x, y))$$